Technical Background Notes for Horizon 2020 Objective ICT-14-2016-2017 Big Data PPP: cross-sectorial and cross-lingual data integration and experimentation

DG CONNECT/G3

CNECT-G3@ec.europa.eu

http://ec.europa.eu/digital-agenda/en/content-and-media/data

http://ec.europa.eu/digital-agenda/en/language-technologies-and-big-data

This document is intended to provide background information and technical commentary on Topic ICT-14 2016 published as part of the 2016-17 Horizon 2020 work programme:

http://ec.europa.eu/research/participants/data/ref/h2020/wp/2016_2017/main/h2020-wp1617-leit-ict_en.pdf

The official work programme text is the only legally binding source of information on the topic. Should any inconsistency between the present explanatory document and the official text be detected it is always to be resolved in favour of the work programme text.

Motivation of the Topic and Scope of this Document

The official text of the work programme specifies that proposals under this objective should be **Innovation Actions**, defined as:

An action primarily consisting of activities directly aiming at producing plans and arrangements or designs for new, altered or improved products, processes or services. For this purpose they may include prototyping, testing, demonstrating, piloting, large-scale product validation and market replication.[1]

This means that research (and publication) activities should **not** be the primary focus of proposals under this objective.

The topic addresses the following challenges:

Specific Challenge: Europe lacks a systematic transfer of knowledge and technology across different sectors and there is an underdeveloped data sharing and linking culture. Traditionally, data has been collected and used for a certain purpose within sectorial "silos", while using data across sectors for offering new services opens new opportunities for solving business and societal challenges. The lack of agreed standards and formats, and the low rates of publishing data assets in machine discoverable formats further hold back data integration. The fact that textual data appears in many languages creates an additional challenge for sharing and linking such data. Finally, there is a lack in Europe of secure environments

[1] http://ec.europa.eu/research/participants/portal/desktop/en/support/reference_terms.html

where researchers and SMEs can test innovative services and product ideas based on open data and business data.

The challenge is to break these barriers and to foster exchange, linking and re-use, as well as to integrate data assets from multiple sectors and across languages and formats. A more specific challenge is to create a stimulating, encouraging and safe environment for experiments where not only data assets but also knowledge and technologies can be shared.

The motivation behind this topic is to create the conditions for value to be created from the interaction of data assets, processes and skills coming from different organizations.

The objective is articulated into two specific types of activities: **data integration** and **data experimentation incubators**.

b) Data Experimentation Incubators

Data experimentation incubator proposals are expected in those domains where there are already European companies willing and ready to make jointly available significant data assets for others to experiment with:

Data experimentation incubators should address big data experimentation in a cross-sectorial, cross lingual and/or cross-border setup. This setup should include access to data in different domains and languages, appropriate computational infrastructure, and open software tools. The incubator should make these available to the experimenters, who are expected to be mainly SMEs, web entrepreneurs and start-ups. Experimentation is to be conducted on horizontal/vertical contributed data pools provided by the incubator.

Examples of this type of collaboration are becoming quite common worldwide in the form of 'data challenges'. Recent examples include:

- http://www.telecomitalia.com/tit/en/bigdatachallenge/contest.html
- http://www.d4d.orange.com/en/Accueil
- http://www.bihapi.pl
- http://www.yelp.com/dataset_challenge
- http://norvigaward.github.io
- http://bigdata.csail.mit.edu/challenge

The strong encouragement given to Small and Medium Enterprises[2] is meant to help creating a European ecosystem of data related expertise, linking large enterprises and their very substantial data assets with agile, smaller, enterprises that have innovative ideas on how to exploit them.

[2] See http://ec.europa.eu/growth/smes/business-friendly-environment/sme-definition/index_en.htm for the official EU definition of a Small and Medium Enterprise

It is precisely to encourage SME participation that proposals under this type of activity are expected to provide all the infrastructure needed for SMEs to concentrate on the execution of their idea: from data assets, to computational infrastructure to software stacks appropriate for the task.

At least half of the experiments should address challenges of industrial importance jointly defined by the data providers, where quantitative performance targets are defined beforehand and results measured against them.

The importance of this requirement is best understood against the background of the Big Data Value Public Private Partnership (PPP) established in October 2014 by the European Commission. In that agreement, the Big Data Value Association[3] representing its data industry members, commits to its industrial members investing €4 of their own resources for every €1 invested made available by the Commission as part of funding available through the Horizon 2020 programme.

The direct consequence of this commitment is that the industrial importance of the challenges defined by the data providers will be evaluated against the evidence that receiving funding for an incubator proposal will cause industrial data providers to invest four-fold **own resources additional to those they would have invested in the absence of funding**. Whilst participation to the Call is completely open to everyone (PPP and non-PPP members), the level of industry investment in activities aimed to scale-up industrial data operations complementary to EC funding is a fundamental impact assessment metric. As the incubators are important multipliers and stimulators of new business, they are especially well placed to enhance the access to financing (e.g venture capital, loans), for example, by match-making and brokerage activities within and around the incubators. It is important that the strategic goals and concrete financial targets (investment levels) are outlined in the proposal. Proposal evaluators will be **specifically instructed** to look for this evidence in submitted proposals. This is why proposals that do not provide this information in a structured, quantitative fashion explicitly tied to a timeline (the PPP commitment is that such industrial investment should take place within the 2016-2020 period) put themselves at an **extreme disadvantage** in what is expected to be a very competitive proposal selection process.

In addition, evidence of the industrial importance of the activities proposed will be linked to explicit specification of performance targets together with a convincing explanation of how such targets are relevant for the business strategy of the affected industrial partners. As a natural complement to those quantitative targets, proposals will need to include a detailed description of the protocol that will be followed to measure if the consortium is advancing at the expected pace towards the targets specified and of the schedule and format in which such measurements will be reported to the Commission. This requirement, on the other hand, does **not** apply for the experiments conducted within the incubator that are more exploratory in nature and not directly related to challenges of industrial importance. Proposal evaluators will be **specifically instructed** to look for this evidence in submitted proposals. Proposals that do

[3] http://bdva.eu/

not provide this information in a structured, quantitative fashion put themselves at an **extreme disadvantage** in what is expected to be a very competitive proposal selection process.

Effective cross-sector and cross-border exchange and re-use of data are key elements in the experiments ecosystem supported by the incubators. Therefore, the incubators are expected to address the technical, linguistic, legal, organisational, and IPR issues, and provide a supported environment for running the experiments.

It is expected that the most innovative ideas will emerge when data assets produced by one type of company (e.g. an agricultural company collecting data about crop quality and yields) are linked, correlated and jointly studied with data assets coming from a different organisation (e.g. a satellite imagery company or a weather data company).

While work funded by this objective is expected to be mostly technical in nature, such work could not even begin if the organisations involved haven't first worked out the legal and security details of such co-operations. Consortium members contributing significant data asset that are commercially or legally sensitive are therefore advised to obtain support from their respective legal departments and high-level decision makers. Similarly, the proposal's risk management plan should specifically discuss measures to be taken should any industrial partners, for whatever reason, withdraw the availability of its data assets. Proposal evaluators will be **specifically instructed** to look for evidence of such support.

To remain flexible on which experiments are carried out and to allow for a fast turn-over of data experimentation activities, the action may involve financial support to third parties, in line with the conditions set out in part K of the General Annexes. The proposal will define the selection process of the experimenters running the data activities for which financial support will be granted (typically in the order of EUR 50 000 – 100 000 per party). At least 70% of the EU funding shall be allocated to this purpose. Experiments are expected to run for a maximum of 6 months, while the incubator should run for a minimum of three years. The proposals are expected to explain how the incubator would become self-sustaining by the end of the funded duration of action.

The incubator model indicated here is a straightforward extension of the model funded in a previous call and visible today in operation in the ODINE incubator[4].

There are two important differences.

The first one is that, while ODINE is mandated to promote the commercial exploitation of open data assets, proposals under this objected are mandated to support experimentation with and exploitation of data assets that are **commercial/industrial** in origin. Open data assets can become part of these experiments, but only as a **supplement** (and **not** a substitute) of commercial/industrial data assets.

The second is that the data incubators that will be selected under this objective will be required to ensure that at least half of the experiments it runs with the participation of third

[4] http://opendataincubator.eu

parties will address industrial challenges with clearly specified performance targets (see above). A direct consequence of this is that the process put in place to select experimenters (typically SMEs) to whom financial support will be granted will have to guarantee this outcome. This in turn implies that the consortium selected to operate the incubator will have to attract applications from experimenters in a manner that makes it very clear what is the nature of the data assets available to the incubator and what are the industrially relevant target performance parameters. Proposal evaluators will be **specifically instructed** to verify that the experimenters' selection process described in proposals has the desired properties.

The impacts expected from the incubator are:

- *At least 100 SMEs and web entrepreneurs, including start-ups, participate in data experimentation incubators;*
- *30% annual increase in the number of Big Data Value use cases supported by the data experimentation incubators;*
- *Substantial increase in the total amount of data made available in the data experimentation incubators including closed data;*
- *Emergence of innovative incubator concepts and business models that allow the incubator to continue operations past the end of the funded duration.*

It is expected that, in order to realistically accomplish this, incubator proposals will request funding in the order of €7M although proposal for different amounts of funding will be considered, if appropriate for the activities proposed.

a) Data Integration

While data experimentation incubator proposals are expected to operate around data assets that are sufficiently well understood in nature and sufficiently well-structured that cross-asset integration can be assumed to be within the ability of a competent software/data engineer, the work programme recognises that there are many other domains where this cannot be assumed:

Data integration activities will address data challenges in cross-domain setups, where similar contributions of data assets will be required by groups of EU industries that are arranged along data value chains (i.e. such that the value extracted by a company in a given industrial sector is greatly increased by the availability and reuse of data produced by other companies in different industrial sectors).

When the business practices of very different business sectors all turn out to need a certain type a data asset (e.g. car companies and smart cities operators both needing constantly updated, high quality maps) it is to be expected that a significant amount of preliminary coordination work should take place to define the nature and format of the data asset of common interest.

The actions will cover the range from informal collaboration to formal specification of standards and will include (but not be limited to) the operation of shared systems of entity

identifiers (so that data about the same entity could be easily assembled from different sources), the definition of agreed data models (so that two companies carrying out the same basic activity would produce data organised in the same way, to the benefit of developers of data analytics tools), support for multilingual data management, data brokerage schemes and the definition of agreed processes to ensure data quality and the protection of commercial confidentiality and personal data. The actions are encouraged to make use of existing data infrastructures and platforms.

A few examples will help explaining what types of activities are expected under this part of the strategic objective.

As an example of a shared system of entity identifiers, one may look at ORCID[5], a non-proprietary alphanumeric code to uniquely identify scientific and other academic authors, currently in use by many commercial publishers; or at the company numbers provided by a business registry such as Companies House[6].

These examples are provided for purposes of illustration only. Any proposal for a shared system of identifiers will be evaluated strictly on the basis of the expected value that the existence of such a system will add to the European economy (either by removing uncertainty and inefficiencies or by creating business opportunities that would be impossible without it).

From this it follows that replicating existing systems of identifiers that already are in widespread use in business processes is unlikely to result in a strong proposal. Similarly, proposals for systems of identifiers that do not yet exist will be convincing only if supported by credible plans for their adoption on the part of a significant percentage of actors within an industry sector or along a supply/business process chain that involves companies from different sectors. Such plans will be credible to the extent that relevant European industrial actors (whether or not formal members of the consortium) are willing to signal their commitment. Describing investment plans that depend on the existence of such systems of identifiers would be a very convincing way to prove commitment.

As an example of a data brokerage scheme, one could consider the e-invoicing scheme recently mandated by law in Italy[7] which, while mandated specifically for those who supply goods or services to public administrations, also makes it easier for invoices exchanged among private parties to be treated as a structured data asset in the respective business processes. Pondering the potential for extending such a scheme across EU Member States[8] also makes it clear why support for multi-lingual data management is strongly expected from proposals addressing this aspect of the objective.

As in the previous case, these examples are provided for purposes of illustration only. Any proposal for data brokerage schemes will be evaluated strictly on the basis of the expected

[5] https://en.wikipedia.org/wiki/ORCID
[6] https://beta.companieshouse.gov.uk/
[7] http://www.fatturapa.gov.it/export/fatturazione/en/normativa/f-1.htm and http://www.fatturapa.gov.it/export/fatturazione/en/normativa/f-2.htm
[8] http://www.peppol.eu/

value that the existence of such schemes will add to the European economy (either by removing uncertainty and inefficiencies or by creating business opportunities that would be impossible without it).

Once again, proposals for such schemes will be evaluated based on the likelihood that they will *actually* be adopted in the short to medium term by significant parts of the European economy. Preliminary work to turn such schemes into actual industry standards, although not required, would be a welcome feature of proposals.

Finally, just as important as data brokerage schemes, will be the specification and testing of business processes that would allow for data assets to become part of economic activities across European companies in a way that makes compliance with data protection regulations[9] and protection of intellectual property rights and commercial confidentiality easier and more efficient to ensure, verify and manage.

To quote again from the official text of the topic, if this is accomplished the expected impact is:

- *Data integration activities will simplify data analytics carried out over datasets independently produced by different companies and shorten time to market for new products and services;*
- *Substantial increase in the number and size of data sets processed and integrated by the data integration activities;*
- *Substantial increase in the number of competitive services provided for integrating data across sectors;*
- *Increase in revenue by 20% (by 2020) generated by European data companies through selling integrated data and data integration services offered.*

As in the case of data incubator proposals, proposals addressing this aspect of the objective will be evaluated in terms of their ability to credibly document how the funding requested will be matched by **additional** investment of European industrial partners in the period 2016-2020 (i.e. investments that would **not** have occurred if the proposal had not been funded. The PPP commitment is to invest 4€ for every 1€ made available by the EC), whether or not listed as formal members of the consortium.

Notice

The information requested for the proper assessment of the work described in a proposal including

1. Data assets made available to the consortium
2. Industrial performance targets for the experiments of industrial importance run in the incubator

[9] http://ec.europa.eu/justice/data-protection/

3. Industrial commitment to additional investments (as required by the terms of the Big Data Value Public Private Partnership agreement)

must be presented in the main body of the description of work (what is known as Part B Section 1) and not in the section (Part B Section 2) describing the members of the consortium.

Appendix: a list of questions that proposals must answer in order to be in scope of objective ICT-14 2016-17 of Horizon 2020

This appendix contains a list of simple questions that a consortium should ask about the proposal to be submitted. If the proposal as submitted does not contain a clear answer to the majority of the relevant questions it places itself at a serious disadvantage in a very competitive selection process (because the evaluators will be specifically instructed to look for the answers to these and other questions)

b) Data Incubators

1. Are the industrial performance targets listed in order to select third party experimenters going to improve an existing industrial/commercial process or to create the conditions for novel industrial/commercial processes?
2. If the incubator experiments are meant to improve existing industrial/commercial processes, please provide **explicit**, **quantitative** data (ideally displayed in a dedicated table, which the evaluators will be instructed to look for and whose absence to note) on:
    a. their cost structure[10]
    b. their current technological constraints/limits
    c. why can the listed performance targets be expected to make European companies more competitive in the time frame foreseen for the work proposed?
3. If the incubator experiments are meant to improve existing industrial/commercial processes, please provide **explicit** and **verifiable** details (ideally displayed in a dedicated table, which the evaluators will be instructed to look for and whose absence to note) on:
    a. How do you plan to measure and report changes in the cost structure of the processes at the end of the experiment (if funded)
    b. How you plan to measure and report changes in the technological constraints by the end of the experiment (if funded)
4. If the incubator experiments are going to use data in order to create the conditions for novel industrial/commercial process, please explain:
    a. The decision process, within each relevant industrial partner, that will determine if the results of a successful experiment will be deployed in the partners actual business operations
    b. How will the relevant partners ensure that they are consistent with the known industrial strategies and existing or planned technological infrastructure of the partners that intend to deploy them in their business

5. Please provide a detailed account of the ownership and user right structure of the data assets to be used and/or produced during the operations of the incubator (if selected

[10] https://en.wikipedia.org/wiki/Market_analysis#Industry_cost_structure

for funding). Provide a detailed account of who will own or have user rights to said data assets after the end of the incubator (if selected for funding).

6. Please describe the industrial strategy and development plans of the commercial companies in your consortium and for each of them (ideally in a dedicated table, which the evaluators will be instructed to look for and whose absence to note) state **explicitly** and in **quantitative**, **verifiable** detail what amount of **own** resources the company intends to invest to leverage the grant received for this proposal, and/or how access to third-party financing (e.g. venture capital) for new business arising from the mini projects is stimulated. If a company has no concrete/verifiable plans to invest additional/own resources, note so explicitly.
7. Using a credible competitiveness analysis framework of your choice[11], provide a credible analysis as to why the changes in industrial/commercial processes resulting from achieving the target industrial performance parameters will make the members of your consortium (and other European industries in the same sector) more competitive **after taking full account of global competition**. For each commercial partner of the consortium provide (in a dedicated table, which the evaluators will be instructed to look for and whose absence to note) a motivated **quantitative** estimate of the resulting increase in market share (together with an equally **quantitative** estimate of the total size of the market by the end of the pilot).
8. Please describe in detail any **legal constraint** that the specific industry sector participating in the incubator imposes on the collection and management of the data assets to be used. Similarly for national legislation that affects consortium members.

a) Data Integrators

1. What evidence do you have that the relevant industrial sectors will benefit from integrating their data assets in the way you are proposing?
    a. What is the business logic behind such a proposal (this requires taking into account the global competitive landscape)?
    b. What is the decision process that will lead to the relevant industrial players to adopt your solution (this requires providing evidence of commitment from the relevant industrial parties, whether or not they are part of the consortium)?
2. What is the estimated cost of the current absence of integration? What do you project the cost of the relevant industrial processes will be if integration is accomplished?
3. Will the integration you propose enable processes that today are impossible or too expensive to pursue?

[11] E.g. https://en.wikipedia.org/wiki/Porter_five_forces_analysis